

# Compression for quantum population coding

Yuxiang Yang<sup>a</sup>, Ge Bai<sup>a</sup>, Giulio Chiribella<sup>a,b</sup> and Masahito Hayashi<sup>c,d</sup>

<sup>a</sup>Department of Computer Science, The University of Hong Kong

<sup>b</sup> Canadian Institute for Advanced Research, CIFAR Program in Quantum Information Science

<sup>c</sup>Graduate School of Mathematics, Nagoya University

<sup>d</sup> Centre for Quantum Technologies, National University of Singapore

Email: yangyx09@gmail.com, bg95@163.com, giulio@hku.hk & masahito@math.nagoya-u.ac.jp

## Abstract

We study the compression of arbitrary parametric families of  $n$  identically prepared finite-dimensional quantum states, in a setting that can be regarded as a quantum analogue of population coding. For a family with  $f$  free parameters, we propose an asymptotically faithful protocol that requires a memory of overall size  $(f/2)\log n$ . Our construction uses a quantum version of local asymptotic normality and, as an intermediate step, solves the problem of the optimal compression of  $n$  identically prepared displaced thermal states. Our protocol achieves the ultimate bound predicted by quantum Shannon theory. In addition, we explore the minimum requirement for quantum memory: On the one hand, the amount of quantum memory used by our protocol can be made arbitrarily small compared to the overall memory cost; on the other hand, any protocol using only classical memory cannot be faithful.

## Index Terms

population coding, compression, quantum system, local asymptotic normality, identically prepared state

## I. INTRODUCTION

Many problems in quantum information theory involve a source that prepares multiple copies of the same quantum state. This is the case, for example, of quantum tomography [1], quantum cloning [2], [3], and quantum state discrimination [4]. The state prepared by the source is generally unknown to the agent who has to carry out the task. Instead, the agent knows that the state belongs to some parametric family of density matrices  $\{\rho_\theta\}_{\theta \in \Theta}$ , with the parameter  $\theta$  varying in the set  $\Theta$ . Also, it is promised that all the particles emitted by the source are independently prepared in the same quantum state  $\rho_\theta$ : when the source is used  $n$  times, it generates  $n$  quantum particles in the tensor product state  $\rho_\theta^{\otimes n}$ .

How much information is contained in the  $n$ -particle state  $\rho_\theta^{\otimes n}$ ? One way to address this question is to quantify the minimum amount of memory needed to store the state. Solving this problem requires an optimization over all possible compression protocols. When the number of copies is large, it is tempting to use a classical protocol, wherein the parameter  $\theta$  is estimated and the estimate is stored in a classical memory. However, this type of storage is generally not faithful, as shown in [5] for several examples of pure state families. In order to achieve a faithful storage, a non-zero amount of quantum memory is generally required. It is important to stress that the problem of storing the  $n$ -copy states  $\{\rho_\theta^{\otimes n}, \theta \in \Theta\}$  in a quantum memory is different from the standard problem of quantum data compression [6], [7], [8]. In our scenario, the mixed state  $\rho_\theta$  is not regarded as the average state of an information source, but, instead, as a physical encoding of the parameter  $\theta$ . The goal of compression is to preserve the encoding of the parameter  $\theta$ , by storing the state  $\rho_\theta^{\otimes n}$  into a memory and retrieving it with high fidelity for all possible values of  $\theta$ . To stress the difference with standard quantum compression, we refer to our scenario as *compression for quantum population coding*. The expression “quantum population coding” refers to the encoding of the parameter  $\theta$  into the many-particle state  $\rho_\theta^{\otimes n}$ . We choose this expression in analogy with (classical) population coding, whereby a parameter is encoded into the population of  $n$  individuals [9]. The typical example of population coding arises in computational neuroscience, where the population consists of neurons and the parameter represents an external stimulus.

The compression for quantum population coding has been studied by Plesch and Bužek [10] in the case where  $\rho_\theta$  is a pure qubit state and no error is tolerated (see also [11] for a prototype experimental implementation). A first extension to mixed states, higher dimensions, and non-zero error was proposed by some of us in [12]. The

protocol therein was proven to be optimal under the assumption that the decoding operation must satisfy a suitable conservation law. Later, a new protocol that reaches the ultimate information-theoretic bound was found for qubit states [13]. The classical version of the problem was addressed in [14]. However, finding the optimal protocol for arbitrary parametric families of quantum states has remained as an open problem so far.

In this paper, we provide the general theory for the compression of  $n$ -tensor product state in a quantum parametric state family. We consider two categories of state families: families of finite-dimensional states and displaced thermal families of infinite-dimensional states. These two categories of state families turn out to be connected by the quantum version of local asymptotic normality (Q-LAN)[15], [16], [17], [18], which reduces  $n$ -tensor product of a finite-dimensional state locally to a displaced thermal state. As the first step, we discuss this kind of compression for the thermal states family, which can be regarded as the quantum extension of the Gaussian distribution. In the next step, employing Q-LAN, we reduce the problem of compressing generic finite-dimensional states to the case of displaced thermal states. Unlike previous works, our protocol does not require any assumption on the symmetry of the state family. In addition, an intriguing feature of this protocol is that the ratio between the size of quantum memory and the size of classical memory can be made arbitrarily close to but not equal to zero. Such a feature is not an incidence but an essence: for identically prepared displaced thermal states and qudit states, we show that any compression protocol using only classical memory must have non-vanishing error.

The rest of the paper is structured as follows. In Section III we study the compression of displaced thermal states. In Section IV we propose a protocol for the compression of identically prepared finite-dimensional states. Optimality of the protocols is proven later in Section VI. In Section V we show that it is necessary to use quantum memory to achieve faithful compression. Finally, we conclude the paper by some discussions in Section VII.

## II. MAIN RESULT.

The main result of our work is the optimal compression of identically prepared quantum states. We consider two major categories of states: finite dimensional (i.e. qudit) states and displaced thermal states. The key question addressed here is how much memory is needed at least to encode the states, in a way that they can be recovered with an error vanishing in the number of input copies. The memory cost essentially depends on the (sub)family from which the states are drawn. For instance, the memory cost for diagonal qudit states (i.e. classical probability distributions) should be less than the cost for general qudit states. As a consequence, we need to define the state subfamilies being considered before stating the main result.

We begin by introducing the parameterization for  $d(<\infty)$ -dimensional non-degenerate quantum systems, whose states can be generated by rotating a fixed state  $\rho_0(\mu)$  with spectrum  $\mu$ , i.e.

$$\rho_\theta = U_\xi \rho_0(\mu) U_\xi^\dagger \quad (1)$$

where

$$U_\xi = \exp \left[ i \left( \sum_{1 \leq j < k \leq d} \frac{\xi_{j,k}^R T_{j,k} + \xi_{j,k}^I T_{k,j}}{\sqrt{\mu_j - \mu_k}} \right) \right] \quad (2)$$

$$T_{j,k} = iE_{j,k} - iE_{k,j} \quad T_{k,j} = E_{j,k} + E_{k,j} \quad (3)$$

is the exponential form of a  $SU(d)$  element,  $E_{j,k}$  is a  $d \times d$  matrix with entry  $(j,k)$  equal to 1 and other entries equal to 0. Therefore, any non-degenerate qudit state can be parameterized as  $\rho_\theta$ , where  $\theta = (\mu, \xi) \in \mathbb{R}^{d^2-1}$  is a vector of parameters with  $\mu = (\mu_1, \mu_2, \dots, \mu_{d-1})$  being its spectrum, ordered as  $\mu_1 > \dots > \mu_{d-1} > 0$ , and  $\xi = (\dots, \xi_{j,k}^R, \xi_{j,k}^I, \dots)$  ( $1 \leq j < k \leq d$ ) being the rotation parameters. A qudit state (sub)family will be denoted as  $\{\rho_\theta^{\otimes n}\}_{\theta \in \Theta}$  with  $\Theta = \Theta_1 \times \Theta_2 \times \dots \times \Theta_{d^2-1}$  where  $\Theta_i$  is the range of the  $i$ -th component of  $\theta$ . For simplicity of discussion, we assume  $\Theta_i = [a_i, a'_i]$  for  $a_i, a'_i \in \mathbb{R}$ . We call a component of  $\mu$  a free *classical parameter* if its range is  $[\mu_i, \mu'_i]$  for  $\mu'_i > \mu_i$  and a component of  $\xi$  a free *quantum parameter* if its range is  $[\xi_i, \xi'_i]$  for  $\xi'_i > \xi_i$ . Otherwise the range set of the parameter contains only one value and the parameter is fixed. We denote by  $f_c$  ( $f_q$ ) the number of free classical (quantum) parameters of the (sub)family.

Next we introduce displaced thermal states, which are a type of states frequently encountered in quantum optics. Displaced thermal states can be regarded as coherent states subject to Gaussian additive noise. A coherent state, describing the state of a laser under ideal conditions, is parametrized as

$$|\alpha\rangle = e^{-\frac{|\alpha|^2}{2}} \sum_{k=0}^{\infty} \frac{\alpha^k}{\sqrt{k!}} |k\rangle \quad \alpha = |\alpha|e^{iT} \quad T \in [0, 2\pi),$$

where  $\{|k\rangle\}$  is the photon number basis. Gaussian additive noise is described by the quantum channel (completely positive trace preserving map)

$$\mathcal{N}_\beta(\rho) = \int d^2\mu \left( \frac{1-\beta}{\pi\beta} \right) e^{-\frac{(1-\beta)|\mu|^2}{\beta}} D_\mu \rho D_\mu^\dagger \quad (4)$$

where  $D_\mu = \exp(\mu\hat{a}^\dagger - \bar{\mu}\hat{a})$  is the displacement operator and  $\beta \in [0, 1)$  is a suitable parameter. When applied to an input coherent state, the Gaussian additive noise outputs the displaced thermal state

$$\rho_{\alpha,\beta} = D_\alpha \rho_\beta^{(\text{thm})} D_\alpha^\dagger \quad \alpha = |\alpha|e^{iT} \quad T \in [0, 2\pi), \quad (5)$$

with

$$\rho_\beta^{(\text{thm})} = \mathcal{N}_\beta(|0\rangle\langle 0|) = (1-\beta) \sum_{i=0}^{\infty} \beta^i |i\rangle\langle i|. \quad (6)$$

We consider a (sub)family of  $n$  identically prepared displaced thermal states, denoted by  $\{\rho_{\alpha,\beta}^{\otimes n}\}_{(\beta,\alpha) \in \Theta}$  with  $\Theta = \Theta_\beta \times \Theta_\alpha$  being the parameter space. Similar as the qudit state case, there are three real parameters for the displaced thermal state family: the thermal parameter  $\beta$ , the strength of the displacement  $|\alpha|$ , and the phase  $T = \arg \alpha$ .  $\beta$  is a classical parameter, specifying the probability distribution of the eigenvalues, while  $|\alpha|$  and  $T$  are quantum parameters, determining the eigenstates of the density matrix. Each of the three parameters is free if its range is  $[a, b]$  for  $a < b$ , otherwise we assume that the range set contains only a single value and the parameter is fixed. Again, we denote by  $f_c$  ( $f_q$ ) the number of free classical (quantum) parameters of the (sub)family.

A compression protocol consists of two components: the encoder, which compresses the input state into a memory, and the decoder, which recovers the state from the memory. A protocol is thus represented by a couple of quantum channels (completely positive trace-preserving linear maps)  $(\mathcal{E}, \mathcal{D})$  characterizing the encoder and the decoder, respectively. We focus on *faithful* compression protocols, whose error vanishes in the large  $n$  limit. As a measure of error, we choose the supremum of the trace distance between the original state and the recovered state  $\mathcal{D} \circ \mathcal{E}(\rho_\theta^{\otimes n})$

$$\varepsilon := \sup_{\theta \in \Theta} \frac{1}{2} \|\rho_\theta^{\otimes n} - \mathcal{D} \circ \mathcal{E}(\rho_\theta^{\otimes n})\|_1. \quad (7)$$

The main result of our work is the following:

**Theorem 1.** *Let  $\{\rho_\theta^{\otimes n}\}_{\theta=(\mu,\xi) \in \Theta}$  be the state (sub)family of  $n$  identical displaced thermal states or non-degenerate qudit states with  $f_c$  free classical parameters and  $f_q$  free quantum parameters. For any  $\delta \in (0, 2/9)$ , the state family can be compressed into  $[(1/2 + \delta)f_c + (1/2)f_q] \log n$  classical bits and  $(f_q\delta) \log n$  qubits with an error  $\varepsilon = O(n^{-\delta/2}) + O(n^{-\kappa(\delta)})$ , where  $\kappa(\delta)$  is the error of Q-LAN [18] given by Eq. (26). The compression is optimal, in the sense that any compression protocol requiring a memory of size  $[(f_c + f_q)/2 - \delta'] \log n$  with  $\delta' > 0$  cannot be faithful.*

Theorem 1 states that to encode each free parameter a memory of size  $(1/2 + \delta) \log n$  is required. When the parameter is classical, the required memory is fully classical; when the parameter is quantum, a quantum memory of  $\delta \log n$  qubits is required. Note that Theorem 1 solves the compression of several important (sub)families:

- The full model family of qudits ( $f_c = d - 1$  and  $f_q = d(d - 1)$ ).
- The classical qudit subfamily ( $f_c = d - 1$  and  $f_q = 0$ ) of  $d$ -dimensional classical probability distributions can be compressed into  $(d/2) \log n$  classical bits, retrieving the result of [14].
- The phase-covariant qudit subfamily ( $f_c = d - 1$  and  $f_q = d(d - 1)/2$ ), where  $\xi_{j,k}^I = 0$  for any  $j, k$ .

### III. COMPRESSION OF DISPLACED THERMAL STATES.

In this section, we consider the compression of identically prepared displaced thermal states  $\rho_{\alpha,\beta}^{\otimes n}$  for 8 different subfamilies, corresponding to cases when  $\beta$ ,  $|\alpha|$  and  $T$  are either free or fixed. The memory costs are summarized in Table I, which matches the statement of Theorem 1. For instance, in Case 5 we have  $f_q = 1$  ( $T$  is a quantum parameter) and  $f_c = 1$  ( $\beta$  is a classical parameter), and therefore the memory costs are  $\delta \log n$  qubits and  $(1 + \delta) \log n$  bits. On the other hand, the errors for all cases satisfy the unified bound

$$\varepsilon = O\left(n^{-\delta/2}\right).$$

TABLE I  
COMPRESSION RATE FOR DIFFERENT STATE FAMILIES. HERE  $\delta > 0$  IS AN ARBITRARY POSITIVE CONSTANT.

Case	displacement $\alpha =  \alpha e^{iT}$	thermal parameter $\beta$	quantum bits	classical bits
0	fixed	fixed	0	0
1	fixed	free	0	$(1/2 + \delta) \log n$
2	free	fixed	$2\delta \log n$	$\log n$
3	free	free	$2\delta \log n$	$(3/2 + \delta) \log n$
4	$T$ free; $ \alpha $ fixed	fixed	$\delta \log n$	$(1/2) \log n$
5	$T$ free; $ \alpha $ fixed	free	$\delta \log n$	$(1 + \delta) \log n$
6	$ \alpha $ free; $T$ fixed	fixed	$\delta \log n$	$(1/2) \log n$
7	$ \alpha $ free; $T$ fixed	free	$\delta \log n$	$(1 + \delta) \log n$

We now show details of the compression protocol for each case. Since Case 0 (all parameters are fixed) is essentially trivial, we start from Case 1, where only  $\beta$  is a free parameter.

#### A. Case 1: free $\beta$ ; fixed $\alpha$ .

An important property for  $n$  identically displaced thermal states is that they are equivalent, up to an unitary transformation, to a product of thermal states where the displacement appears only on one mode, namely

$$U_{BS} \rho_{\alpha,\beta}^{\otimes n} U_{BS}^\dagger = \rho_{\sqrt{n}\alpha,\beta} \otimes \left( \rho_{\beta}^{(\text{thm})} \right)^{\otimes (n-1)}, \quad (8)$$

where  $U_{BS}$  is a unitary operator, which can be physically implemented through a suitable arrangement of beam splitters [19], [20]. Using this property, we can construct a protocol that separately compresses the displaced thermal state  $\rho_{\sqrt{n}\alpha,\beta}$  and the  $n - 1$  thermal modes. For the compression of the thermal modes, we have the following lemma.

**Lemma 1** (Compression of identically prepared thermal states). *For any  $x > 0$ , there exists a protocol  $(\mathcal{E}_{n,x}^{(\text{thm})}, \mathcal{D}_{n,x}^{(\text{thm})})$  compressing  $n$  copies of a thermal state  $\rho_{\beta}^{(\text{thm})}$  into  $(1/2 + x) \log n$  classical bits with error*

$$\varepsilon_{\text{thm}} = O(n^{-x}). \quad (9)$$

The proof of the above lemma can be found in Appendix. We emphasize that no quantum memory is required to encode thermal states. For any  $\delta > 0$ , we can then construct the compression protocol as follows:

- *Encoder.*

- 1) First perform on each input copy the displacement operation  $\mathcal{D}_{-\alpha}$ , defined by

$$\mathcal{D}_{\mu}(\cdot) := D_{\mu} \cdot D_{\mu}^\dagger$$

where  $D_{\mu} = \exp(\mu \hat{a}^\dagger - \bar{\mu} \hat{a})$  is the displacement operator. The displacement operation transforms each input copy  $\rho_{\alpha,\beta}$  into the thermal state  $\rho_{\beta}^{(\text{thm})}$ .

- 2) Then apply the thermal state encoder  $\mathcal{E}_{n,\delta}^{(\text{thm})}$  on the  $n$ -mode state and the outcome is encoded in a classical memory.

- *Decoder.*

- 1) First use the thermal state decoder  $\mathcal{D}_{n,\delta}^{(\text{thm})}$  to recover the  $n$  copies of the thermal state  $\rho_{\beta}^{(\text{thm})}$  from the classical memory.

- 2) Perform the displacement operation  $\mathcal{D}_\alpha$  on each mode.

Note that the input state can be regarded as the state of  $n$  optical modes with each mode in a displaced thermal state, and thus the compression protocol can be regarded as a sequence of operations on the  $n$ -mode system. We shall use the term “mode” throughout this section.

The above protocol requires only  $(1/2 + \delta)\log n$  bits for any  $\delta > 0$  by Lemma 1. The error is also the same as in Lemma 1, for the displacement operations are unitary.

### B. Case 2: unknown $\alpha$ , fixed $\beta$ .

Next we study the more involved case when the displacement  $\alpha$  is unknown. Let us first look at the idea how to compress before going into details of the compression protocol. To save as much quantum memory as possible, we may consider to estimate  $\alpha$  and store the outcome in a classical memory. However, any estimate of  $\alpha$  comes at the price of disturbing the input state, and, as shown later in Section V, the distortion caused by measurements on all input copies is so large that the state cannot be recovered faithfully. Instead of full measurements, here we adopt a strategy of measuring only a small portion of the input copies so that we still get an estimate of  $\alpha$ , while the distortion can still be recovered. We then perform coherent quantum protocols based on the information gained by the estimation. Such an idea of gently testing the input state and then performing coherent compression is key to this work, which allows us to convert the memory cost from quantum to classical as much as possible.

For any  $\delta > 0$ , the protocol runs as follows:

- *Preprocessing.* A preprocessing procedure is needed in order to store the estimate of  $\alpha$ : Divide the range of  $\alpha$  into  $n$  zones, each labeled by a point  $\hat{\alpha}_i$  in it, so that  $|\alpha' - \alpha''| = O(n^{-1/2})$  (note that  $\alpha$  is complex) for any  $\alpha', \alpha''$  in the same zone.
- *Encoder.*

- 1) First perform the unitary channel  $\mathcal{U}_{\text{BS}}(\cdot) = U_{\text{BS}} \cdot U_{\text{BS}}^\dagger$  on the input state, where  $U_{\text{BS}}$  is the unitary defined by Eq. (8). The output state has the form  $\rho_{\sqrt{n}\alpha, \beta} \otimes \left(\rho_\beta^{(\text{thm})}\right)^{\otimes (n-1)}$ .
- 2) Send the first and the last mode through a beam splitter with transitivity  $n^{-\delta}$ . The  $n$ -mode state is now  $\rho_{\sqrt{n-n^{1-\delta}}\alpha, \beta} \otimes \left(\rho_\beta^{(\text{thm})}\right)^{\otimes (n-2)} \otimes \rho_{\sqrt{n^{1-\delta}}\alpha, \beta}$ .
- 3) Estimate  $\alpha$  by the heterodyne measurement  $\{d^2\alpha'|\alpha'\rangle\langle\alpha'|\}$  on the last mode, which yields an estimate  $\hat{\alpha} = \alpha'/\sqrt{n^{1-\delta}}$  with the probability distribution

$$Q(\hat{\alpha}|\alpha) = (1 - \beta) \exp[-(1 - \beta)n^{1-\delta}|\hat{\alpha} - \alpha|^2]. \quad (10)$$

Encode the label of the zone  $\hat{\alpha}^*$  containing  $\hat{\alpha}$  in a classical memory.

- 4) Displace the first mode with  $\mathcal{D}_{-\sqrt{n-n^{1-\delta}}\hat{\alpha}^*}$ .
- 5) The state of the first mode then goes through a truncation in the photon number basis, described by the channel  $\mathcal{P}_{n^{2\delta}}$  defined as

$$\mathcal{P}_{n^{2\delta}}(\rho) = P_{n^{2\delta}}\rho P_{n^{2\delta}} + (1 - \text{Tr}[P_{n^{2\delta}}\rho])|0\rangle\langle 0| \quad (11)$$

where

$$P_{n^{2\delta}} = \sum_{m=0}^{n^{2\delta}} |m\rangle\langle m|. \quad (12)$$

The output state on the first mode is encoded in a quantum memory.

- *Decoder.*

- 1) First read  $\hat{\alpha}^*$  and perform the displacement operation  $\mathcal{D}_{\sqrt{n-n^{1-\delta}}\hat{\alpha}^*}$  on the state of the quantum memory.
- 2) To recover the distortion caused by the estimation of  $\alpha$ , a quantum amplifier [21]  $\mathcal{A}^{\gamma_n}$  with  $\gamma_n = 1/(1 - n^{-\delta})$  is applied to the output state. A quantum amplifier is a device that increases the intensity of quantum light while preserving its phase information, defined by the following quantum channel

$$\mathcal{A}^\gamma(\rho) = \text{Tr}_B \left[ e^{\cosh^{-1}(\sqrt{\gamma})(a^\dagger b^\dagger - ab)} (\rho \otimes |0\rangle\langle 0|_B) e^{\cosh^{-1}(\sqrt{\gamma})(ab - a^\dagger b^\dagger)} \right] \quad (13)$$

where  $a$  and  $b$  are the annihilation operators of the input mode and the ancillary mode  $B$ , respectively.

- 3) Prepare the other  $(n-1)$  modes in the thermal state  $\rho_{\beta}^{(\text{thm})}$ .
- 4) Perform on all the  $n$  modes the inverse of the unitary channel  $\mathcal{U}_{\text{BS}}$

$$\mathcal{U}_{\text{BS}}^{-1}(\rho) = U_{\text{BS}}^{\dagger} \rho U_{\text{BS}}. \quad (14)$$

The memory cost consists of two parts: the cost of encoding the (rounded) value  $\hat{\alpha}^*$  of the estimate which is  $\log n$  bits and the cost of encoding the first mode (in a displaced thermal state) which is  $2\delta \log n$  qubits. Overall, the protocol requires  $2\delta \log n$  qubits and  $\log n$  classical bits.

On the other hand, the error of the protocol can be split into three terms as

$$\begin{aligned} \varepsilon \leq & \int_{|\hat{\alpha}-\alpha| > n^{-1/2+3\delta/4}} Q(d\hat{\alpha}|\alpha) \\ & + \frac{1}{2} \sup_{\alpha, \beta} \sup_{\hat{\alpha}^*: |\hat{\alpha}^*-\alpha| \leq n^{-1/2+3\delta/4}} \left\{ \left\| \mathcal{A}^{\gamma_n}(\rho_{\sqrt{n-n^{1-\delta}}\alpha, \beta}) - \rho_{\sqrt{n}\alpha, \beta} \right\|_1 + \left\| \mathcal{P}_{n^{2\delta}}(\rho_{\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*), \beta}) - \rho_{\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*), \beta} \right\|_1 \right\}. \end{aligned}$$

Now we bound them one by one. First, from Eq. (10) we get

$$\int_{|\hat{\alpha}-\alpha| > n^{-1/2+3\delta/4}} Q(d\hat{\alpha}|\alpha) = \int_{|\hat{\alpha}-\alpha| > n^{-1/2+3\delta/4}} (1-\beta) \exp[-(1-\beta)n^{1-\delta}|\hat{\alpha}-\alpha|^2] = e^{-\Omega(n^{\delta/2})}.$$

Second, explicit calculation shows that, when the input state is a displaced thermal state  $\rho_{\alpha, \beta}$ , the output state of the amplifier is

$$\mathcal{A}^{\gamma}(\rho_{\alpha, \beta}) = \rho_{\alpha, \beta'} \quad \beta' = \frac{\beta\gamma + 4(1-\beta)\gamma(\gamma-1)}{\gamma + (1-\beta)(\gamma-1)}. \quad (15)$$

We thus have

$$\frac{1}{2} \sup_{\alpha, \beta} \sup_{\hat{\alpha}^*: |\hat{\alpha}^*-\alpha| \leq n^{-1/2+3\delta/4}} \left\| \mathcal{A}^{\gamma_n}(\rho_{\sqrt{n-n^{1-\delta}}\alpha, \beta}) - \rho_{\sqrt{n}\alpha, \beta} \right\|_1 = O(n^{-\delta})$$

having used Eq. (19). Finally, the error of  $\mathcal{P}_{n^{2\delta}}$  can be bounded using the following lemma, whose proof can be found in Appendix.

**Lemma 2** (Photon number truncation of displaced thermal states.). *Define the channel  $\mathcal{P}_N$  as*

$$\mathcal{P}_N(\rho) = P_N \rho P_N + (1 - \text{Tr}[P_N \rho])|0\rangle\langle 0| \quad (16)$$

where

$$P_N = \sum_{k=0}^N |k\rangle\langle k|. \quad (17)$$

When  $N = \Omega(|\alpha|^{2+x})$ ,  $\mathcal{P}_N$  satisfies

$$\varepsilon(\rho_{\alpha, \beta}) = \frac{1}{2} \left\| \mathcal{P}_N(\rho_{\alpha, \beta}) - \rho_{\alpha, \beta} \right\|_1 = \beta^{\Omega(N^{x/8})} + e^{-\Omega(N^{x/4})} \quad (18)$$

for any  $0 \leq \beta < 1$ .

In our case we are using the projector  $P_{n^{2\delta}}$  in Eq. (12), and the displacement is  $\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*)$ . Since  $(n-n^{1-\delta})|\alpha-\hat{\alpha}^*|^2 = O(n^{3\delta/2})$ , by Lemma 2 we obtain the bound

$$\frac{1}{2} \sup_{\alpha, \beta} \sup_{\hat{\alpha}^*: |\hat{\alpha}^*-\alpha| \leq n^{-1/2+3\delta/4}} \left\| \mathcal{P}_{n^{2\delta}}(\rho_{\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*), \beta}) - \rho_{\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*), \beta} \right\|_1 = \beta^{\Omega(n^{\delta/12})} + e^{-\Omega(n^{\delta/6})}$$

Summarizing the above bounds for each term of the error, we get

$$\varepsilon = O(n^{-\delta}).$$



### C. Case 3: unknown $\alpha$ , unknown $\beta$ .

Case 3 can be treated in a similar way as Case 2, except that we also have to estimate and encode  $\beta$ . Luckily, unlike  $\alpha$ , the thermal parameter  $\beta$  can be estimated freely (i.e. without distortion of the input state), and thus its estimation strategy is simpler.

In detail, for any  $\delta > 0$  we can then construct the protocol for displaced thermal states with unknown  $\beta$  and unknown  $\alpha$  as follows:

- *Preprocessing.* Divide the range of  $\alpha$  into  $n$  zones, each labeled by a point  $\hat{\alpha}_i$  in it, so that  $|\alpha' - \alpha''| = O(n^{-1/2})$  for any  $\alpha', \alpha''$  in the same zone.
- *Encoder.*
  - 1) First perform the unitary channel  $\mathcal{U}_{\text{BS}}(\cdot) = U_{\text{BS}} \cdot U_{\text{BS}}^\dagger$  on the input state, where  $U_{\text{BS}}$  is the unitary defined by Eq. (8). The output state has the form  $\rho_{\sqrt{n}\alpha, \beta} \otimes (\rho_{\beta}^{(\text{thm})})^{\otimes(n-1)}$ .
  - 2) Estimate  $\beta$  with the von Neumann measurement of the photon number on the  $n-1$  copies of  $\rho_{\beta}^{(\text{thm})}$  and denote by  $\hat{\beta}$  the maximum likelihood estimate of  $\beta$ . Note that these copies will not be disturbed since they are diagonal in the photon number basis.
  - 3) Next, send the first and the last mode through a beam splitter with transitivity  $n^{-\delta}$ . The  $n$ -mode state is now  $\rho_{\sqrt{n-n^{1-\delta}}\alpha, \beta} \otimes (\rho_{\beta}^{(\text{thm})})^{\otimes(n-2)} \otimes \rho_{\sqrt{n^{1-\delta}}\alpha, \beta}$ .
  - 4) Estimate  $\alpha$  by the heterodyne measurement  $\{d^2\alpha'|\alpha'\rangle\langle\alpha'|\}$  on the last mode, which yields an estimate  $\hat{\alpha} = \alpha'/\sqrt{n^{1-\delta}}$  with the probability distribution  $\mathcal{Q}(\hat{\alpha}|\alpha)$  (10). Encode the label of the zone  $\hat{\alpha}^*$  containing  $\hat{\alpha}$  in a classical memory.
  - 5) Displace the first mode with  $\mathcal{D}_{-\sqrt{n-n^{1-\delta}}\hat{\alpha}^*}$ .
  - 6) Prepare the  $n$ -th mode in the thermal state  $\rho_{\hat{\beta}}^{(\text{thm})}$ . The  $n$ -mode state is now  $\rho_{\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*), \beta} \otimes (\rho_{\beta}^{(\text{thm})})^{\otimes(n-2)} \otimes \rho_{\hat{\beta}}^{(\text{thm})}$ .
  - 7) The state of the first mode then goes through a truncation in the photon number basis, described by the channel  $\mathcal{P}_{n^{2\delta}}$  defined by Eq. (16). The output state on the first mode is encoded in a quantum memory.
  - 8) Meanwhile, use the thermal state encoder  $\mathcal{E}_{n-1, \delta}^{(\text{thm})}$  (see Lemma 1) to compress the remaining  $n-1$  modes and encode the output state in a classical memory.
- *Decoder.*
  - 1) First read  $\hat{\alpha}^*$  and perform the displacement  $\mathcal{D}_{\sqrt{n-n^{1-\delta}}\hat{\alpha}^*}$  on the state of the quantum memory.
  - 2) Apply a quantum amplifier  $\mathcal{A}^{\gamma_n}$  with  $\gamma_n = 1/(1-n^{-\delta})$  to the state.
  - 3) Then use the thermal state decoder  $\mathcal{D}_{n-1, \delta}^{(\text{thm})}$  to recover the other  $(n-1)$  modes in the thermal state  $\rho_{\beta}^{(\text{thm})}$  from the memory.
  - 4) Finally, perform the inverse of the unitary channel  $\mathcal{U}_{\text{BS}}$ .

The memory cost consists of three parts: the cost of encoding the (rounded) value  $\hat{\alpha}^*$  of the estimate which is  $\log n$  bits, the cost of encoding the first mode (displaced thermal state) which is  $2\delta \log n$  qubits, and the cost of encoding the other modes (thermal states) which is  $(1/2 + \delta) \log n$  bits. Overall, the protocol requires  $2\delta \log n$  qubits and  $(3/2 + \delta) \log n$  classical bits.

On the other hand, the error of the protocol can be split into several terms as

$$\begin{aligned} \varepsilon \leq & \frac{1}{2} \sup_{\beta} \left\| \mathcal{D}_{n, \delta}^{(\text{thm})} \circ \mathcal{E}_{n, \delta}^{(\text{thm})} \left[ \int d\hat{\beta} P(d\hat{\beta}|\beta) \rho_{\hat{\beta}}^{(\text{thm})} \otimes (\rho_{\beta}^{(\text{thm})})^{\otimes(n-1)} \right] - (\rho_{\beta}^{(\text{thm})})^{\otimes n} \right\|_1 \\ & + \frac{1}{2} \sup_{\alpha, \beta} \sup_{\hat{\alpha}^*: |\hat{\alpha}^* - \alpha| \leq n^{-1/2+3\delta/4}} \left\{ \left\| \mathcal{P}_N^{(\text{num})}(\rho_{\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*), \beta}) - \rho_{\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*), \beta} \right\|_1 + \left\| \mathcal{A}^{\gamma_n}(\rho_{\sqrt{n-n^{1-\delta}}\alpha, \beta}) - \rho_{\sqrt{n}\alpha, \beta} \right\|_1 \right\} \\ & + \int_{|\hat{\alpha} - \alpha| > n^{-1/2+3\delta/4}} \mathcal{Q}(d\hat{\alpha}|\alpha) \end{aligned}$$

where  $P(d\hat{\beta}|\beta)$  denotes the probability distribution of the estimate for the input value  $\beta$ . The latter three terms are the errors of estimation and compression of the displaced thermal state  $\rho_{\sqrt{n}\alpha, \beta}$  of the first mode, which have been

bounded in the previous subsection as  $O(n^{-\delta})$ . We now bound the first term  $\epsilon_\beta$ , which is the error of compressing and estimating thermal states.  $\epsilon_\beta$  can be further split into two terms as

$$\epsilon_\beta \leq \frac{1}{2} \sup_{\beta} \left\{ \left\| \mathcal{D}_{n,\delta}^{(\text{thm})} \circ \mathcal{E}_{n,\delta}^{(\text{thm})} \left[ \left( \rho_{\beta}^{(\text{thm})} \right)^{\otimes n} \right] - \left( \rho_{\beta}^{(\text{thm})} \right)^{\otimes n} \right\|_1 + \left\| \int P(d\hat{\beta}|\beta) \rho_{\hat{\beta}}^{(\text{thm})} - \rho_{\beta}^{(\text{thm})} \right\|_1 \right\}.$$

On the right hand side of the inequality, the first error term is bounded by Lemma 1, while the second error term is bounded by the following property of the maximum likelihood estimate [22]

$$\int_{|\hat{\beta}-\beta| \geq l/\sqrt{nF_\beta}} P(d\hat{\beta}|\beta) \leq \text{erfc} \left( \frac{l}{\sqrt{2}} \right)$$

where  $F_\beta = (\beta^2 + 1)/[\beta(1 - \beta)^3]$  is the Fisher information of  $\beta$  and  $\text{erfc}(x) := (2/\pi) \int_x^\infty e^{-s^2} ds$  is the complementary error function.  $\epsilon_\beta$  can thus be bounded as

$$\begin{aligned} \epsilon_\beta &\leq O(n^{-\delta}) + \frac{1}{2} \sup_{\beta} \sup_{\hat{\beta}: |\hat{\beta}-\beta| \leq n^{-(1+\delta)/2}} \left\| \rho_{\hat{\beta}}^{(\text{thm})} - \rho_{\beta}^{(\text{thm})} \right\|_1 + e^{-\Omega(n^\delta)} \\ &= O(n^{-\delta}) \end{aligned}$$

having used the tail property of complementary error function and the property

$$\left\| \rho_{\alpha}^{(\text{thm})} - \rho_{\beta}^{(\text{thm})} \right\|_1 = \frac{2|\alpha - \beta|}{(1 - \alpha)^2} + O(|\alpha - \beta|^2). \quad (19)$$

Summarizing the above bounds for each term of the error, we get

$$\epsilon = O(n^{-\delta}).$$

*D. Case 4: fixed  $|\alpha|$ , free  $T$ , fixed  $\beta$  and Case 6: free  $|\alpha|$ , fixed  $T$ , fixed  $\beta$ .*

In Case 4 and Case 6, the displacement  $\alpha$  is partially known. Such a knowledge allows us to reduce the amount of memory. Since the protocols for these two cases are very similar. The protocol for Case 4 runs as in the following:

- *Preprocessing.* Divide the range of  $T$  into  $n^{-1/2}$  zones, each labeled by a point  $\hat{T}_i$  in it, so that  $|T' - T''| = O(n^{-1/2})$  for any  $T', T''$  in the same zone.
- *Encoder.*
  - 1) First perform the unitary channel  $\mathcal{U}_{\text{BS}}(\cdot)$  on the input state to transform it into  $\rho_{\sqrt{n}\alpha, \beta} \otimes \left( \rho_{\beta}^{(\text{thm})} \right)^{\otimes (n-1)}$ .
  - 2) Next, send the first and the last mode through a beam splitter with transitivity  $n^{-\delta/2}$ . The  $n$ -mode state is now  $\rho_{\sqrt{n-n^{1-\delta/2}}\alpha, \beta} \otimes \left( \rho_{\beta}^{(\text{thm})} \right)^{\otimes (n-2)} \otimes \rho_{\sqrt{n^{1-\delta/2}}\alpha, \beta}$ .
  - 3) Estimate  $T$  by the heterodyne measurement  $d^2\alpha'|\alpha'\rangle\langle\alpha'|$  on the last mode, which yields an estimate  $\hat{T}$  which is the phase of  $\alpha'$ . Encode the label of the zone  $\hat{T}^*$  containing  $\hat{T}$  in a classical memory.
  - 4) Displace the first mode with  $\mathcal{D}_{-\sqrt{n-n^{1-\delta/2}}\hat{\alpha}^*}$  with  $\hat{\alpha}^* := |\alpha|e^{i\hat{T}^*}$ .
  - 5) Then send the state of the first mode through a truncation channel  $\mathcal{P}_{n^\delta}$  defined in (16) and encode the output state in a quantum memory.
- *Decoder.*
  - 1) First read  $\hat{\alpha}^*$  and perform the displacement  $\mathcal{D}_{\sqrt{n-n^{1-\delta/2}}\hat{\alpha}^*}$  on the state of the quantum memory.
  - 2) Then apply the quantum amplifier  $\mathcal{A}_n^{\gamma_n}$  with  $\gamma_n = 1/(1 - n^{-\delta/2})$ .
  - 3) Prepare the other  $(n-1)$  modes in the thermal state  $\rho_{\beta}^{(\text{thm})}$ .
  - 4) Finally, perform  $\mathcal{U}_{\text{BS}}^{-1}$  on the output of  $\mathcal{D}_{n-1}^{(\text{thm})}$  and the quantum memory.

The protocol for Case 6 works in the same way except that  $|\alpha|$  is estimated instead of  $T$ . For both cases the memory cost consists of three parts: the cost of encoding the (rounded) value  $\hat{\alpha}^*$  of the estimate which is  $(1/2) \log n$  bits, the cost of encoding the first mode (displaced thermal state) which is  $\delta \log n$  qubits, and the cost of encoding the other modes (thermal states) which is  $(1/2 + \delta) \log n$  bits. Overall, the protocol requires  $\delta \log n$  qubits and  $(1 + 2\delta) \log n$  classical bits. The error can be bounded as previous as

$$\epsilon = O(n^{-\delta/2}).$$



E. Case 5: fixed  $|\alpha|$ , unknown  $T$ , unknown  $\beta$  and Case 7: unknown  $|\alpha|$ , fixed  $T$ , unknown  $\beta$ .

Case 5 and Case 7 can be treated in the same way as Case 4 and Case 6, adding additional steps to estimate and encode  $\beta$ . The protocol for Case 5 runs as follows:

- *Preprocessing.* Divide the range of  $T$  into  $n^{-1/2}$  zones, each labeled by a point  $\hat{T}_i$  in it, so that  $|T' - T''| = O(n^{-1/2})$  for any  $T', T''$  in the same zone.
- *Encoder.*
  - 1) First perform the unitary channel  $\mathcal{U}_{BS}(\cdot)$  on the input state to transform it into  $\rho_{\sqrt{n}\alpha, \beta} \otimes \left(\rho_{\beta}^{(\text{thm})}\right)^{\otimes(n-1)}$ .
  - 2) Estimate  $\beta$  with the von Neumann measurement of the photon number on the  $n-1$  copies of  $\rho_{\beta}^{(\text{thm})}$ . Denote by  $\hat{\beta}$  the maximum likelihood estimate of  $\beta$ .
  - 3) Next, send the first and the last mode through a beam splitter with transitivity  $n^{-\delta/2}$ . The  $n$ -mode state is now  $\rho_{\sqrt{n-n^{-\delta/2}}\alpha, \beta} \otimes \left(\rho_{\beta}^{(\text{thm})}\right)^{\otimes(n-2)} \otimes \rho_{\sqrt{n^{-1-\delta/2}}\alpha, \beta}$ .
  - 4) Estimate  $T$  by the heterodyne measurement  $d^2\alpha'|\alpha'\rangle\langle\alpha'|$  on the last mode, which yields an estimate  $\hat{T}$  which is the phase of  $\alpha'$ . Encode the label of the zone  $\hat{T}^*$  containing  $\hat{T}$  in a classical memory.
  - 5) Displace the first mode with  $\mathcal{D}_{-\sqrt{n-n^{-\delta/2}}\hat{\alpha}^*}$  with  $\hat{\alpha}^* := |\alpha|e^{i\hat{T}^*}$ .
  - 6) Prepare the  $n$ -th mode in the thermal state  $\rho_{\beta}^{(\text{thm})}$ . The  $n$ -mode state is now  $\rho_{\sqrt{n-n^{-\delta/2}}(\alpha-\hat{\alpha}^*), \beta} \otimes \left(\rho_{\beta}^{(\text{thm})}\right)^{\otimes(n-2)} \otimes \rho_{\beta}^{(\text{thm})}$ .
  - 7) Then send the state of the first mode through a truncation channel  $\mathcal{P}_{n\delta}$  defined in (16) and encode the output state in a quantum memory.
  - 8) Meanwhile, use the thermal state encoder  $\mathcal{E}_{n-1, \delta}^{(\text{thm})}$  (see Lemma 1) to compress the remaining  $n-1$  modes and encode the output state in a classical memory.
- *Decoder.*
  - 1) First read  $\hat{\alpha}^*$  and perform the displacement  $\mathcal{D}_{\sqrt{n-n^{-\delta/2}}\hat{\alpha}^*}$  on the state of the quantum memory.
  - 2) Then apply the quantum amplifier  $\mathcal{A}^{\gamma_n}$  with  $\gamma_n = 1/(1-n^{-\delta/2})$ .
  - 3) Meanwhile, use the thermal state decoder  $\mathcal{D}_{n-1, \delta}^{(\text{thm})}$  to recover the other  $(n-1)$  modes in the thermal state  $\rho_{\beta}^{(\text{thm})}$  from the memory.
  - 4) Finally, perform  $\mathcal{U}_{BS}^{-1}$  on the output of  $\mathcal{D}_{n-1}^{(\text{thm})}$  and the quantum memory.

The protocol for Case 7 works in the same way except that  $|\alpha|$  is estimated instead of  $T$ . For both cases the memory cost consists of three parts: the cost of encoding the (rounded) value  $\hat{\alpha}^*$  of the estimate which is  $(1/2)\log n$  bits, the cost of encoding the first mode (displaced thermal state) which is  $\delta\log n$  qubits, and the cost of encoding the other modes (thermal states) which is  $(1/2 + \delta)\log n$  bits. Overall, the protocol requires  $\delta\log n$  qubits and  $(1 + 2\delta)\log n$  classical bits. The error can be bounded as previous as

$$\varepsilon = O\left(n^{-\delta/2}\right).$$

#### IV. COMPRESSION OF IDENTICALLY PREPARED FINITE DIMENSIONAL SYSTEMS.

In this section, we study the compression of  $d(<\infty)$ -dimensional non-degenerate quantum systems, based on the results on displaced thermal states and quantum local asymptotic normality. We show that, just as for displaced thermal states, each free parameter of a qudit subfamily requires  $(1/2 + \delta)\log n$  memory for any  $\delta > 0$ . Details of the compression protocol are introduced in the following.

##### A. The compression protocol

To construct a compression protocol, we need the following techniques:

- *Quantum local asymptotic normality (Q-LAN).* The quantum version of local asymptotic normality has been derived in several different contexts [16], [17], [18]. Here we use the result of [18], which states that  $n$  identical copies of a qudit state can be approximated by a classical-quantum Gaussian state in a sufficiently

small neighborhood of a point  $\theta_0 \in \Theta$  for large  $n$ . Explicitly, for a fixed point  $\theta_0 = (\mu_0, \xi_0)$  and for a fixed  $x \in (0, 1)$ , we define the following neighborhood

$$\Theta_{n,x}(\theta_0) = \{\theta = \theta_0 + \delta\theta/\sqrt{n}, \mid \|\delta\theta\|_\infty \leq n^{\frac{x}{2}}\}, \quad (20)$$

where  $\|\delta\theta\|_\infty$  is the max vector norm defined as  $\|\delta\theta\|_\infty := \max_i (\delta\theta)_i$ . Q-LAN states that every  $n$ -fold product state  $\rho_\theta^{\otimes n}$  with  $\theta$  in the neighborhood  $\Theta_{n,x}(\theta_0)$  can be approximated by a classical-quantum Gaussian state. Specifically, the stated  $\rho_\theta^{\otimes n}$  is approximated by the Gaussian state

$$G_{n,\theta} = N_{\delta\mu, I_{\mu_0}} \bigotimes_{j < k} \rho_{\alpha_{j,k}, \beta_{j,k}}, \quad (21)$$

where  $N_{\delta\mu, I_{\mu_0}}$  is the multivariate normal distribution with mean  $\delta\mu$  and covariance matrix  $I_{\mu_0}$  (equal to the inverse of the quantum Fisher information of the eigenvalues  $\mu$ , evaluated at  $\mu_0$ ) and  $\rho_{\alpha_{j,k}, \beta_{j,k}}$  is the displaced thermal state defined as

$$\rho_{\alpha_{j,k}, \beta_{j,k}} = D_{\alpha_{j,k}} \rho_{\beta_{j,k}}^{(\text{thm})} D_{\alpha_{j,k}}^\dagger \quad (22)$$

$$\alpha_{j,k} = \frac{\delta\xi_{j,k}^R + i\delta\xi_{j,k}^I}{2\sqrt{(\mu_0)_j - (\mu_0)_k}} \quad \beta_{j,k} = \frac{(\mu_0)_k}{(\mu_0)_j}. \quad (23)$$

where  $(\mu_0)_j$  and  $(\mu_0)_k$  are components of  $\mu_0$ . The approximation is physically implemented by two quantum channels  $\mathcal{T}_{\theta_0}^{(n)}$  and  $\mathcal{S}_{\theta_0}^{(n)}$ , satisfying the conditions

$$\sup_{\theta \in \Theta_{n,x}(\theta_0, c)} \left\| \mathcal{T}_{\theta_0}^{(n)}(\rho_\theta^{\otimes n}) - G_{n,\theta} \right\|_1 = O\left(n^{-\kappa(x)}\right) \quad (24)$$

$$\sup_{\theta \in \Theta_{n,x}(\theta_0, c)} \left\| \rho_\theta^{\otimes n} - \mathcal{S}_{\theta_0}^{(n)}(G_{n,\theta}) \right\|_1 = O\left(n^{-\kappa(x)}\right), \quad (25)$$

where  $\|\cdot\|_1$  denotes the trace norm and

$$\kappa(x) = \min \left\{ \frac{1-z-\eta}{2}, \frac{1-x-2y}{2} - y, \frac{2-9\eta}{24} \right\} \quad (26)$$

under the constraints  $1 > z > 1+x/2$ ,  $y > 0$ ,  $\eta > 0$  and  $\eta > x-y$ . We note that  $\kappa(x) > 0$  when  $x \in [0, 2/9]$ .

- *Quantum state tomography.* State tomography is an important technique of quantum information processing which is used to determine the density matrix of an unknown quantum state. The role of tomography here is to provide a rough estimate of  $\theta_0$  so that we can apply Q-LAN. The tomography protocol used here [23] gives an estimate  $\rho_{\hat{\theta}}$  of a qudit state  $\rho_\theta$  with confidence

$$\text{Prob} \left[ \frac{1}{2} \|\rho_\theta - \rho_{\hat{\theta}}\|_1 \leq \varepsilon \right] \geq 1 - (n+1)^{3d^2} e^{-n\varepsilon^2} \quad (27)$$

using  $n$  copies of the state.

For any  $\delta \in (0, 2/9)$ , our compression protocol runs as follows:

- *Encode the input state into memories.* We break the encoding of  $\rho_\theta^{\otimes n}$  into four steps:
  - 1) *Tomography.* First take out  $n^{1-\delta/2}$  copies of  $\rho_\theta$  for quantum tomography. In this way, one obtains a neighborhood  $\Theta_{n, 2\delta/3}(\theta_0)$  (note that  $2\delta/3$  is not the only choice) that contains the state with confidence approaching one in the large  $n$  limit. To encode the outcome of the tomography, the parameter space  $\Theta$  is discretized into a lattice of  $n^{(f_c+f_q)/2}$  points, each point corresponding to an outcome of tomography. The point that is closest to  $\theta_0$  is encoded in a classical memory.
  - 2) *Q-LAN.* After the tomography step,  $n - n^{1-\delta/2}$  copies remain for compression. Define

$$\gamma_n = 1/(1 - n^{-\delta/2}). \quad (28)$$

so that the number of remaining copies is  $n/\gamma_n$ . The  $n/\gamma_n$  copies are sent through the channel  $\mathcal{T}_{\theta_0}^{(n/\gamma_n)}$  (24) which outputs the Gaussian state  $G_{(n/\gamma_n), \theta}$  defined by Eq. (21).

- 3) *Amplification.* Next, the Gaussian state is amplified to compensate the loss of input copies. The state  $\rho_{\alpha_{j,k}, \beta_{j,k}}$  of each quantum mode is amplified by  $\mathcal{A}^{\gamma_n}$ . The Gaussian distribution on the classical register is rescaled by a constant factor:

$$|u\rangle \rightarrow |\sqrt{\gamma_n}u\rangle,$$

where  $\{|u\rangle\}$  stands for the Cartesian basis of the classical register. The whole amplification progress is described by the channel  $\mathcal{A}_{\theta_0}^{(n/\gamma_n) \rightarrow n}$ , whose action on any product state is

$$\mathcal{A}_{\theta_0}^{(n/\gamma_n) \rightarrow n}(\rho \otimes_{j < k} \sigma_{j,k}) = \sum_u \langle u | \rho | u \rangle |\sqrt{\gamma_n}u\rangle \langle \sqrt{\gamma_n}u| \bigotimes_{j < k} \mathcal{A}^{\gamma_n}(\sigma_{j,k}). \quad (29)$$

where  $u$  is summed over the basis of the classical register.

- 4) *Gaussian state compression.* Each quantum mode of the amplified Gaussian state is then truncated by  $\mathcal{P}_{n^\delta}$  (16). The output state is then stored in a quantum memory. The state of the classical mode is compressed by an channel  $\mathcal{P}_c$  that truncates the state into a  $O(n^\delta)$ -hypercube centered around the mean of the Gaussian. Explicitly, we have

$$\mathcal{P}_c(\rho) = \sum_{\|u\|_\infty \leq n^\delta} \langle u | \rho | u \rangle |u\rangle \langle u| + \left[ 1 - \sum_{\|u'\|_\infty \leq n^\delta} \langle u' | \rho | u' \rangle \right] |0\rangle \langle 0|. \quad (30)$$

The whole process is described by the channel

$$\mathcal{P}_{\theta_0}^{(n)} = \mathcal{P}_c \bigotimes_{j < k} \mathcal{P}_{n^\delta}. \quad (31)$$

- *Recover the original state.* The state can be decompressed from the memory by sending the state of the hybrid memory through the channel  $\mathcal{S}_{\theta_0}^{(n)}$  (25), which can be constructed by consulting the outcome of tomography.

## B. Error analysis.

Here we bound the error of the protocol. Using the triangle inequality of trace distance, we split the overall error into four terms

$$\varepsilon \leq \varepsilon_{\text{tomo}} + \varepsilon_{\text{amp}} + \varepsilon_G + \varepsilon_{\text{LAN}}, \quad (32)$$

where

$$\varepsilon_{\text{tomo}} = \mathbf{Prob} [\theta \notin \Theta_{n, 2\delta/3}(\theta_0)] \quad (33)$$

$$\varepsilon_{\text{amp}} = \frac{1}{2} \sup_{\theta_0} \sup_{\theta \in \Theta_{n, 2\delta/3}(\theta_0)} \left\| \mathcal{A}_{\theta_0}^{(n/\gamma_n) \rightarrow n}(G_{n/\gamma_n, \theta}) - G_{n, \theta} \right\|_1 \quad (34)$$

$$\varepsilon_G = \frac{1}{2} \sup_{\theta_0} \sup_{\theta \in \Theta_{n, 2\delta/3}(\theta_0)} \left\| \mathcal{P}_{\theta_0}^{(n)}(G_{n, \theta}) - G_{n, \theta} \right\|_1 \quad (35)$$

$$\varepsilon_{\text{Q-LAN}} = \frac{1}{2} \sup_{\theta_0} \sup_{\theta \in \Theta_{n, 2\delta/3}(\theta_0)} \left\{ \left\| \mathcal{T}_{\theta_0}^{(n/\gamma_n)}(\rho_{\theta}^{\otimes (n/\gamma_n)}) - G_{n/\gamma_n, \theta} \right\|_1 + \left\| \rho_{\theta}^{\otimes n} - \mathcal{S}_{\theta_0}^{(n)}(G_{n, \theta}) \right\|_1 \right\} \quad (36)$$

are the error terms of tomography, amplification, truncation, and Q-LAN, respectively.

We first briefly review the process of tomography and then bound its error. Using  $n^{1-\delta/2}$  copies of  $\rho_{\theta}$ , an estimate  $\hat{\theta}$  of  $\theta$  is obtained. The estimate  $\hat{\theta}$  is then encoded as a point in the lattice  $\mathbf{L} := \{\theta \in \Theta \mid |\theta_i| = z_i/(2\sqrt{n}), z_i \in \mathbb{N} \forall i\}$ . Naturally, we encode the coordinate of the lattice point  $\theta_0$  that is closest to  $\hat{\theta}$ . Namely that we choose

$$\theta_0 := \underset{\theta' \in \mathbf{L}}{\operatorname{argmin}} \|\hat{\theta} - \theta'\|_\infty. \quad (37)$$

Recall that  $\|\theta\|_\infty := \max_i(\theta)_i$ . We now bound the error. By definition of the neighborhood  $\Theta_{n, 2\delta/3}(\theta_0)$  (20), we have

$$\varepsilon_{\text{tomo}} = \mathbf{Prob} \left[ \|\theta - \theta_0\|_\infty > n^{-1/2+\delta/3} \right] \quad (38)$$

$$\leq \mathbf{Prob} \left[ \|\theta - \hat{\theta}\|_\infty + \|\hat{\theta} - \theta_0\|_\infty > n^{-1/2+\delta/3} \right] \quad (39)$$

$$\leq \mathbf{Prob} \left[ \|\theta - \hat{\theta}\|_\infty > n^{-1/2+\delta/3}(1 - O(n^{-\delta/3})) \right]. \quad (40)$$

The first inequality comes from triangle inequality and the second inequality is derived by noticing  $\min_{\theta \neq \theta' \in \mathcal{L}} \|\theta - \theta'\|_\infty = (1/2)n^{-1/2}$ , which implies that  $\|\hat{\theta} - \theta_0\|_\infty \leq (1/2)n^{-1/2}$ . To further bound the error, we notice that the parameter space  $\Theta$  allows for the Euclidean expansion of trace distance, namely that there exists a constant  $C$  such that  $\|\rho_\theta - \rho_{\theta'}\|_1 = C\|\theta - \theta'\|_\infty + O(\|\theta - \theta'\|_\infty^2)$ . And thus we have

$$\epsilon_{\text{tomo}} \leq \mathbf{Prob} \left[ \frac{1}{2} \|\rho_\theta - \rho_{\hat{\theta}}\|_1 > (C/4)n^{-1/2+\delta/3} \right]. \quad (41)$$

Substituting  $\epsilon$  with  $(C/4)n^{-1/2+\delta/3}$  and  $n$  with  $n^{1-\delta/2}$  in Eq. (27) we have

$$\epsilon_{\text{tomo}} = n^{-\Omega(n^{\delta/6})}. \quad (42)$$

Next we look at the error of amplification, which can be further split into two terms: the term of classical mode amplification and the term of quantum mode amplification. We first analyze the classical term, which comes from the rescaling of the Gaussian distribution. This operation shifts the center of the normal distribution from  $\sqrt{\gamma_n^{-1}}\delta\mu$  (the classical part of  $G_{n/\gamma_n, \theta}$ ) to  $\delta\mu$ , which coincides with that of  $G_{n, \theta}$ , while it also deforms the covariance matrix from  $I_{\mu_0}$  to  $\gamma_n I_{\mu_0}$ . As a result, we consider the difference of the following distributions:  $N_{\delta\mu, I_{\mu_0}}$  and  $N_{\delta\mu, \gamma_n I_{\mu_0}}$ . As they have the same center, we may translate them both to the origin. The error is:

$$\epsilon_{\text{classical}} = \int_{\mathbb{R}^d} \left| |2\pi I_{\mu_0}|^{-1/2} e^{-\frac{1}{2}\mathbf{x}^T I_{\mu_0}^{-1} \mathbf{x}} - |2\pi \gamma_n I_{\mu_0}|^{-1/2} e^{-\frac{1}{2\gamma_n}\mathbf{x}^T I_{\mu_0}^{-1} \mathbf{x}} \right| d\mathbf{x} \quad (43)$$

$$\leq |2\pi I_{\mu_0}|^{-1/2} \int_{\mathbb{R}^d} \left| e^{-\frac{1}{2}\mathbf{x}^T I_{\mu_0}^{-1} \mathbf{x}} - e^{-\frac{1}{2\gamma_n}\mathbf{x}^T I_{\mu_0}^{-1} \mathbf{x}} \right| d\mathbf{x} + 2\pi n^{-\delta/2} \int_{\mathbb{R}^d} e^{-\frac{1}{2\gamma_n}\mathbf{x}^T I_{\mu_0}^{-1} \mathbf{x}} d\mathbf{x} \quad (44)$$

$$\leq O(n^{-\delta/2}) \int_{\mathbb{R}^d} \mathbf{x}^T I_{\mu_0}^{-1} \mathbf{x} e^{-\frac{1}{2}\mathbf{x}^T I_{\mu_0}^{-1} \mathbf{x}} d\mathbf{x} \quad (45)$$

$$= O(n^{-\delta/2}) \quad (46)$$

Now we check the quantum term. On the quantum register, the amplifier acts independently on each mode as the displaced thermal state amplifier defined by Eq. (13). From a similar calculation as Eq. (15), we get that

$$\epsilon_{\text{quantum}} \leq \frac{1}{2} \sum_{j < k} \left\| \mathcal{A}^{\gamma_n} (\rho_{\alpha_{j,k}, \beta_{j,k}}) - \rho_{\alpha_{j,k}, \beta_{j,k}} \right\|_1, \quad (47)$$

where the error of each quantum amplifier is

$$\frac{1}{2} \left\| \mathcal{A}^{\gamma_n} (\rho_{\alpha_{j,k}, \beta_{j,k}}) - \rho_{\alpha_{j,k}, \beta_{j,k}} \right\|_1 = O(n^{-\delta/2}). \quad (48)$$

Therefore, we conclude that the amplification error scales at most as

$$\epsilon_{\text{amp}} \leq \epsilon_{\text{classical}} + \epsilon_{\text{quantum}} = O(n^{-\delta/2}). \quad (49)$$

Let us now consider the term for the Gaussian state compression, which can be expressed as

$$\epsilon_G \leq \frac{1}{2} \left\| \mathcal{P}_c (N_{\delta\mu, I_{\mu_0}}) - N_{\delta\mu, I_{\mu_0}} \right\|_1 + \frac{1}{2} \sum_{j < k} \left\| \mathcal{P}_{n^\delta} (\rho_{\alpha_{j,k}, \beta_{j,k}}) - \rho_{\alpha_{j,k}, \beta_{j,k}} \right\|_1 \quad (50)$$

For the classical part, noticing that  $\|\delta\mu\|_\infty \leq n^{\delta/3}$ , we have

$$\left\| \mathcal{P}_c (N_{\delta\mu, I_{\mu_0}}) - N_{\delta\mu, I_{\mu_0}} \right\|_1 \leq \sum_{\|u\|_\infty > n^\delta} N_{\delta\mu, I_{\mu_0}}(u) \quad (51)$$

$$\leq \sum_{\|u - \delta\mu\|_\infty > n^\delta - n^{\delta/3}} N_{\delta\mu, I_{\mu_0}}(u) \quad (52)$$

$$= e^{-\Omega(n^{2\delta})} \quad (53)$$

where  $N_{\delta\mu, I_{\mu_0}}(u)$  denotes the probability distribution function. For each of the quantum modes, employing Lemma 2, with  $N$  substituted by  $n^\delta$  and  $|\alpha_{j,k}| = O(n^{\delta/3})$ , we have

$$\frac{1}{2} \left\| \mathcal{P}_{j,k} (\rho_{\alpha_{j,k}, \beta_{j,k}}) - \rho_{\alpha_{j,k}, \beta_{j,k}} \right\|_1 = \beta_{j,k}^{\Omega(n^{\delta/8})} + e^{-\Omega(n^{\delta/4})}. \quad (54)$$

Substituting Eqs. (53) and (54) into Eq. (50), we have

$$\epsilon_G = \max_{j < k} \left[ \frac{(\mu_0)_k}{(\mu_0)_j} \right]^{\Omega(n^{\delta/8})} + e^{-\Omega(n^{\delta/4})}. \quad (55)$$

Finally, we note that the error of the Q-LAN approximation, corresponding to the errors generated by the transformations between the input state and its Gaussian state approximation, is given by Eqs. (24) and (25) as

$$\epsilon_{\text{Q-LAN}} = O\left(n^{-\kappa(\delta)}\right). \quad (56)$$

Summarizing the above bounds (42), (49), (55), (56) on each of the error terms, we conclude that the protocol generates an error which scales at most

$$\epsilon = O\left(n^{-\kappa(\delta)}\right) + O\left(n^{-\delta/2}\right) \quad (57)$$

### C. Memory cost.

There are three sources of memory cost: a classical memory of  $[(f_c + f_q)/2] \log n$  bits for the tomography outcome, a classical memory of  $f_c \delta \log n$  bits for the classical part of the Gaussian state and a quantum memory of  $f_q \delta \log n$  qubits for the quantum part of the Gaussian state. It takes  $(1/2 + \delta) \log n$  bits to encode a classical free parameter and  $(1/2) \log n$  bits plus  $\delta \log n$  qubits to encode a quantum free parameter.

From the above discussion we can see that the ratio between the quantum memory cost and the classical memory cost is

$$R_{q/c} = \frac{\delta f_q}{(1/2 + \delta) f_c + (1/2) f_q}, \quad (58)$$

which can be made close to zero when  $\delta$  is set close to zero. This result shows that the quantum memory cost can be made arbitrarily small compared to the classical memory cost.

## V. NECESSITY OF A QUANTUM MEMORY.

From the previous discussion, we know that the quantum memory cost can be made small compared to the classical memory cost. It then natural to ask whether the quantum memory is necessary at all or it is enough to perform full tomography on the input state and to use only a classical memory. Here we show that compression with a fully classical memory cannot be faithful, for both displaced thermal states and qudit states. A quick glance sees that, if the memory is fully classical, the concatenation of the encoder and the decoder would be a measure-and-prepare channel. The problem is then turned into *the error benchmark* problem, i.e. the problem of finding the minimal worst case error for measure-and-prepare transmissions of a family of states. The error benchmark  $\epsilon_c$  for a family of states  $\{\rho_x\}_{x \in \mathcal{X}}$  is defined as

$$\epsilon_c(\{\rho_x\}_{x \in \mathcal{X}}) := \inf_{\mathcal{C}^{(\text{MP})}} \sup_{\mathbf{x} \in \mathcal{X}} \frac{1}{2} \left\| \mathcal{C}^{(\text{MP})}(\rho_x) - \rho_x \right\|_1 \quad (59)$$

where the infimum is taken over all measure-and-prepare quantum channels  $\mathcal{C}^{(\text{MP})}$ .

Let us start from the  $n$  identically prepared displaced thermal state case. From existing results [24], we already know that the error benchmark for displaced thermal states is strictly positive

$$\epsilon_c(\{\rho_{\alpha, \beta}\}_{\alpha \in \mathbb{C}}) = (2 - \beta)^{-i_\beta - 1} - \beta^{i_\beta + 1} > 0 \quad (60)$$

where  $i_\beta$  is the integer part of  $-\log(2 - \beta)/\log \beta(2 - \beta)$ . The idea is to show that any compression protocol for displaced thermal states must at least have an error  $\epsilon_c(\{\rho_{\alpha, \beta}\}_{\alpha \in \mathbb{C}})$  in the asymptotic limit of large  $n$ .

Consider any compression protocol  $(\mathcal{E}_n, \mathcal{D}_n)$  for  $\{\rho_{\alpha, \beta}^{\otimes n}\}_{(\beta, \alpha) \in \Theta}$  using only classical memory. Note that we assume without loss of generality that the point  $(\beta, \alpha = 0)$  and its neighborhood are contained in the interior of  $\Theta$  for some  $\beta$ , otherwise the same argument applies up to a displacement operation. We can then fix this  $\beta$ , and construct a measure-and-prepare channel  $\mathcal{C}_n^{(\text{MP})}$  for the displaced thermal state family  $\{\rho_{\alpha, \beta}\}_{\alpha \in \mathbb{C}}$  with the following operational description:

- 1) For any input  $\rho_{\alpha, \beta}$ , first use a beam splitter to split the state into  $\rho_{\sqrt{1-t^2}\alpha, \beta} \otimes \rho_{t\alpha, \beta}$  for  $t = n^{-1/8}$ .

- 2) Apply the heterodyne measurement to the second system whose state is  $\rho_{t\alpha,\beta}$ , obtaining the estimate  $\hat{\alpha}$  of  $\alpha$  with the probability distribution  $Q(\hat{\alpha}|\alpha) = (1-\beta)\exp[-(1-\beta)t^2|\hat{\alpha}-\alpha|^2]$ .
- 3) Apply the displacement  $\mathcal{D}_{-\sqrt{1-t^2}\hat{\alpha}}$  to the first system. The resultant state is  $\rho_{\sqrt{1-t^2}(\alpha-\hat{\alpha}),\beta}$ .
- 4) Use beam splitters to transform the state into  $\rho_{\alpha',\beta}^{\otimes n}$ , where  $\alpha' = \sqrt{1-t^2}(\alpha-\hat{\alpha})/\sqrt{n}$ .
- 5) Apply  $\mathcal{D}_n \circ \mathcal{E}_n$ .
- 6) Use beam splitters to transform the state back into  $\rho_{\sqrt{1-t^2}(\alpha-\hat{\alpha}),\beta}$  on the first mode.
- 7) Apply the displacement  $\mathcal{D}_{\sqrt{1-t^2}\hat{\alpha}}$  to the first system. The resultant state is (approximately)  $\rho_{\sqrt{1-t^2}\alpha,\beta}$ .
- 8) Apply the amplifier  $\mathcal{A}^{\gamma}$  (13) with  $\gamma = 1/\sqrt{1-t^2}$ .

The error of the above protocol can be bounded as

$$\begin{aligned} \frac{1}{2} \|C_n^{(\text{MP})}(\rho_{\alpha,\beta}) - \rho_{\alpha,\beta}\|_1 &\leq \int_{\sqrt{1-t^2}|\hat{\alpha}-\alpha| > n^{1/4}} Q(d\hat{\alpha}|\alpha) + \frac{1}{2} \|\mathcal{A}^{\gamma}(\rho_{\sqrt{1-t^2}\alpha,\beta}) - \rho_{\alpha,\beta}\|_1 \\ &\quad + \frac{1}{2} \sup_{\alpha': \sqrt{1-t^2}|\hat{\alpha}-\alpha| \leq n^{1/4}} \|\mathcal{D}_n \circ \mathcal{E}_n(\rho_{\alpha',\beta}^{\otimes n}) - \rho_{\alpha',\beta}^{\otimes n}\|_1 \\ &\leq e^{-\Omega(n^{1/4})} + O(n^{-1/4}) + \frac{1}{2} \sup_{\alpha': \sqrt{1-t^2}|\hat{\alpha}-\alpha| \leq n^{1/4}} \|\mathcal{D}_n \circ \mathcal{E}_n(\rho_{\alpha',\beta}^{\otimes n}) - \rho_{\alpha',\beta}^{\otimes n}\|_1. \end{aligned}$$

For large enough  $n$  the set  $\{(\beta, \alpha') \mid \alpha' = \sqrt{1-t^2}(\alpha-\hat{\alpha})/\sqrt{n}, \sqrt{1-t^2}|\hat{\alpha}-\alpha| \leq n^{1/4}\} = \{(\beta, \alpha') \mid |\alpha'| \leq n^{-1/4}\}$  is contained in  $\Theta$ , and thus we have

$$\frac{1}{2} \|C_n^{(\text{MP})}(\rho_{\alpha,\beta}) - \rho_{\alpha,\beta}\|_1 \leq e^{-\Omega(n^{1/4})} + O(n^{-1/4}) + \frac{1}{2} \sup_{\alpha'} \|\mathcal{D}_n \circ \mathcal{E}_n(\rho_{\alpha',\beta}^{\otimes n}) - \rho_{\alpha',\beta}^{\otimes n}\|_1.$$

Taking the limit  $n \rightarrow \infty$  and using Eq. (60) we have

$$\begin{aligned} \frac{1}{2} \lim_{n \rightarrow \infty} \sup_{\alpha'} \|\mathcal{D}_n \circ \mathcal{E}_n(\rho_{\alpha',\beta}^{\otimes n}) - \rho_{\alpha',\beta}^{\otimes n}\|_1 &\geq \lim_{n \rightarrow \infty} \frac{1}{2} \|C_n^{(\text{MP})}(\rho_{\alpha,\beta}) - \rho_{\alpha,\beta}\|_1 \\ &\geq (2-\beta)^{-i_\beta-1} - \beta^{i_\beta+1} > 0. \end{aligned}$$

This proves that  $(\mathcal{E}_n, \mathcal{D}_n)$  is not faithful, and thus displaced thermal states cannot be compressed faithfully using only a classical memory.

The same idea can also be used for identically prepared qudit states thanks to Q-LAN. Namely that we can bound the performance of compression protocols for qudits limited to classical memories with the benchmark (60), using Q-LAN to interconvert between qudits and displaced thermal states.

Consider any compression protocol  $(\mathcal{E}_n, \mathcal{D}_n)$  for the qudit family  $\{\rho_\theta^{\otimes n}\}_{\theta \in \Theta}$  (note that the parametrization is the same as in Eq. 2) using only a classical memory. We can lower bound its error by constructing a measure-and-prepare  $\mathcal{M}_n^{(\text{MP})}$  for the displaced thermal state family  $\{\rho_{\alpha,\beta}\}_{\alpha \in \mathbb{C}}$ , which uses  $(\mathcal{E}_n, \mathcal{D}_n)$  as a subroutine. Choose  $\theta_0 = (\mu_0, \xi_0)$  such that its neighborhood is in  $\Theta$ , and then choose  $\beta = (\mu_0)_{k_0}/(\mu_0)_{j_0}$  for some  $j_0 < k_0$ .  $\mathcal{M}_n^{(\text{MP})}$  consists in the following steps:

- 1) For any input  $\rho_{\alpha,\beta}$ , first use a beam splitter to split the state into  $\rho_{\sqrt{1-t^2}\alpha,\beta} \otimes \rho_{t\alpha,\beta}$  for  $t = n^{-1/11}$ .
- 2) Apply the heterodyne measurement to the second system whose state is  $\rho_{t\alpha,\beta}$ , obtaining the estimate  $\hat{\alpha}$  of  $\alpha$  with the probability distribution  $Q(\hat{\alpha}|\alpha) = (1-\beta)\exp[-(1-\beta)t^2|\hat{\alpha}-\alpha|^2]$ .
- 3) Apply the displacement  $\mathcal{D}_{-\sqrt{1-t^2}\hat{\alpha}}$  to the first system. The resultant state is  $\rho_{\sqrt{1-t^2}(\alpha-\hat{\alpha}),\beta}$ .
- 4) Construct the following classical-quantum Gaussian state similar to Eq. (21) with  $\delta\theta = 0$ :

$$G = N_{0,I_{\mu_0}} \bigotimes_{j < k} \rho_{j,k}$$

except that  $\rho_{j_0,k_0}$  is replaced by the resultant state in the previous step.  $\rho_{j,k} = \rho_{(\mu_0)_k/(\mu_0)_j}^{(\text{thm})}$  for  $(j,k) \neq (j_0,k_0)$ .

- 5) Apply the inverse Q-LAN transformation  $\mathcal{S}_{\theta_0}^{(n)}$  as defined in Eq. (25).
- 6) Apply  $\mathcal{D}_n \circ \mathcal{E}_n$ .
- 7) Apply the Q-LAN transformation  $\mathcal{T}_{\theta_0}^{(n)}$  as defined in Eq. (24).
- 8) Retrieve the mode indexed with  $(j_0, k_0)$ .
- 9) Apply the displacement  $\mathcal{D}_{\sqrt{1-t^2}\hat{\alpha}}$  to the system. The resultant state is (approximately)  $\rho_{\sqrt{1-t^2}\alpha,\beta}$ .



10) Apply the amplifier  $\mathcal{A}_t^\gamma$  (13) with  $\gamma_t = 1/\sqrt{1-t^2}$ .

Now we argue that the error of  $\mathcal{M}_n^{(\text{MP})}$  equals to the error of  $(\mathcal{D}_n, \mathcal{E}_n)$ , up to terms vanishing in the limit of large  $n$ . We only need to justify that the error of Q-LAN transformations vanishes, since the error analysis for the other parts have already been done in the  $n$  identically prepared displaced thermal state case.

Define the following neighborhood as in Eq. (20) with  $x = 1/5$ :

$$\Theta_{n,x}(\theta_0) = \left\{ \theta = \theta_0 + \delta\theta/\sqrt{n}, \mid \|\delta\theta\| \leq n^{\frac{1}{10}} \right\}, \quad (61)$$

We can obviously choose  $n$  large enough such that  $\Theta_{n,x}(\theta_0) \subseteq \Theta$ . We then address that, with high probability, the resultant state  $\rho_{\sqrt{1-t^2}(\alpha-\hat{\alpha}),\beta}$  in Step 3) makes the Gaussian state  $G$  actually in the image of Q-LAN transformation  $\mathcal{T}_{\theta_0}^{(n)}$  with valid input, so that  $\mathcal{S}_{\theta_0}^{(n)}(G)$  recovers approximately  $\rho_{\theta}^{\otimes n}$  with some  $\theta \in \Theta_{n,x}(\theta_0) \subseteq \Theta$ . According to the form of  $G$  and Eq. (22), we have  $\theta = \theta_0 + \delta\theta/\sqrt{n}$  with  $\delta\theta = (\delta\mu, \delta\xi)$ , where  $\delta\mu = 0$ ,  $\delta\xi_{j_0,k_0} = 2\sqrt{(\mu_0)_j - (\mu_0)_k}\sqrt{1-t^2}(\alpha - \hat{\alpha})$ , and  $\delta\xi_{j,k} = 0$  for  $(j,k) \neq (j_0,k_0)$ . The probability that  $\theta \notin \Theta_{n,x}(\theta_0)$  is

$$\begin{aligned} \mathbf{Prob}[\theta \notin \Theta_{n,x}(\theta_0)] &\leq \mathbf{Prob}\left[2\sqrt{(\mu_0)_j - (\mu_0)_k}\sqrt{1-t^2}|\alpha - \hat{\alpha}| > n^{\frac{1}{10}}\right] \\ &= e^{-\Omega(n^{1/55})}. \end{aligned}$$

using the tail of Gaussian distribution where the variance of  $(\hat{\alpha} - \alpha)$  is  $O(n^{2/11})$ . In the limit  $n \rightarrow \infty$ , using Eq. (60) we have

$$\frac{1}{2} \limsup_{n \rightarrow \infty} \sup_{\alpha'} \|\mathcal{D}_n \circ \mathcal{E}_n(\rho_{\alpha',\beta}^{\otimes n}) - \rho_{\alpha',\beta}^{\otimes n}\|_1 \geq (2 - \beta)^{-i_\beta - 1} - \beta^{i_\beta + 1} > 0.$$

This proves that displaced thermal states cannot be compressed faithfully using only a classical memory.

## VI. OPTIMALITY OF THE COMPRESSION

The optimality proof of the compression protocol requires us to quantify the capacity of ensembles of  $\rho_{\theta}^{\otimes n}$ , i.e. to find an ensemble  $\{p_{\theta}, \rho_{\theta}^{\otimes n}\}$  with large enough Holevo information [25]  $\chi := S(\sum_{\theta} p_{\theta} \rho_{\theta}^{\otimes n}) - \sum_{\theta} p_{\theta} S(\rho_{\theta}^{\otimes n})$  [ $S(\cdot)$  denotes the von Neumann entropy]. Since any protocol requires a memory of size  $\log n_{\text{enc}} \approx \log \chi$  to encode the ensemble faithfully, we can get a good lower bound on  $n_{\text{enc}}$ . We notice that the parameter space  $\Theta$  allows for the Euclidean expansion of trace distance, namely that there exists a constant  $C$  such that  $\|\rho_{\theta} - \rho_{\theta'}\|_1 = C\|\theta - \theta'\|_{\infty} + O(\|\theta - \theta'\|_{\infty}^2)$ . Now we define a mesh  $M$  on the parameter space  $\Theta$ , on which points can be almost perfectly distinguished from each other:

$$M = \{\theta \in \Theta \mid |(\theta - \theta_0)_i| = z_i \cdot \log n / \sqrt{n}, z_i \in \mathbb{N} \forall i\} \quad (62)$$

where  $\theta_0 \in \Theta$  is a fixed point. The mesh  $M$  is so defined that

$$\varepsilon_{\min} := \min_{\theta \neq \theta' \in M} \frac{1}{2} \|\rho_{\theta} - \rho_{\theta'}\|_1 = \frac{C \log n}{2\sqrt{n}} + O\left(\frac{\log^2 n}{n}\right). \quad (63)$$

On the other hand, the number of points contained in the mesh satisfies

$$|M| \geq T_{\Theta} \left[ \frac{\sqrt{n}}{\log n} \right]^{f_c + f_q}, \quad (64)$$

where  $T_{\Theta} > 0$  is independent of  $n$ . We consider an ensemble of states with parameters uniformly distributed in  $M$ , defined as

$$C_n = \left\{ \rho_{\theta}^{\otimes n}, \sum_{\theta_0 \in M} |M|^{-1} \delta_{\theta_0, \theta} \right\}. \quad (65)$$

We now show that the ensemble  $C_n$  can be used to transmit  $\log |M|$  bits of classical information almost perfectly. First of all, we notice that the states in  $C_n$  are nearly perfectly distinguishable from each other. For qudit state subfamilies, define the following POVM elements distinguishing states as

$$\tilde{M}_{\theta} = \int_{\frac{1}{2} \|\sigma - \rho_{\theta}\|_1 \leq \frac{\varepsilon_{\min}}{2}} d\sigma M_{\sigma} \quad (66)$$

where  $M_\sigma$  is the qudit tomography POVM element defined in Eq. (9) of [23], which the property that

$$\int_{\frac{1}{2}\|\sigma-\rho\|_1 \leq \varepsilon/2} d\sigma \operatorname{Tr}[M_\sigma \rho^{\otimes n}] \geq 1 - (n+1)^{3d^2} e^{-n\varepsilon^2}. \quad (67)$$

(For displaced thermal state families, a similar result holds for the heterodyne measurement of  $\alpha$  and maximum likelihood estimation of  $\beta$  [19].) The probability of error when this POVM is used to distinguish states in  $C_n$  can be bounded as

$$P_e(\theta) = \int_{\frac{1}{2}\|\rho_{\theta'} - \sigma\|_1 > \frac{\varepsilon_{\min}}{2}} d\sigma \operatorname{Tr}[\rho_\theta^{\otimes n} M_\sigma] \quad (68)$$

$$\leq (n+1)^{3d^2} e^{-\frac{C \log^2 n}{4}} \quad (69)$$

$$\leq n^{-\frac{C \log n}{8}}. \quad (70)$$

A communication protocol for conveying the message  $\theta \in M$  is to transmit the quantum state  $\rho_m^{\otimes n}$  and then to get the recovered message  $\hat{\theta}$  using the POVM  $\{\tilde{M}_\theta\}$ . By Fano's inequality [26], the Holevo information of the ensemble can be lower bounded as

$$\chi(C_n) \geq I(\Theta : \hat{\Theta}) \quad (71)$$

$$\geq (1 - P_e) \log |M| - h(P_e) \quad (72)$$

$$= \frac{f_c + f_q}{2} \log n - (f_c + f_q) \log \log n + o(1). \quad (73)$$

where  $S(\rho)$  denotes the von Neumann entropy,  $h(x) = -x \log x$  ( $h(0) := 0$ ), and

$$P_e = \frac{\int d\theta P_e(\theta)}{\int d\theta} \leq n^{-\frac{C \log n}{8}} \quad (74)$$

is the average error probability.

Finally, using the bound for quantum compression [27] we conclude that the number of qubits required is lower bounded as

$$n_{\text{enc}} \geq \frac{f_c + f_q}{2} \log n - (f_c + f_q) \log \log n + h(\varepsilon) + h(1 - \varepsilon) + O(\varepsilon \log n). \quad (75)$$

For  $\varepsilon$  vanishing faster than the inverse of  $\log n$ , the leading order of the compression cost is  $[(f_c + f_q)/2] \log n$ , which justifies the optimality of our protocol.

## VII. CONCLUSION

In this work we have solved the problem of compressing identically prepared states of finite-dimensional quantum systems and identically prepared displaced thermal states. We showed that the size of the required memory is proportional to the number of free parameters of the state. Especially, we note that finite-dimensional states can be stored in an almost-classical memory. Our compression protocol can be applied to study a quantum version of the problem of population coding, in the simple scenario where the population consists of identical and independently prepared quantum particles. However, in a real microscopic system, particles typically have correlation. An interesting and challenging problem, left to future research, is the extension of our results to the study of correlated quantum manybody systems.

### Acknowledgments

This work is supported by the Canadian Institute for Advanced Research (CIFAR), by the Hong Kong Research Grant Council through Grant No. 17326616, by National Science Foundation of China through Grant No. 11675136, and by the HKU Seed Funding for Basic Research, and the John Templeton Foundation. Y. Y. is supported by a Hong Kong and China Gas Scholarship. MH was supported in part by a MEXT Grant-in-Aid for Scientific Research (B) No. 16KT0017, the Okawa Research Grant and Kayamori Foundation of Informational Science Advancement.

## REFERENCES

- [1] Konrad Banaszek, Marcus Cramer, and David Gross. Focus on quantum tomography. *New Journal of Physics*, 15(12):125020, 2013.
- [2] Valerio Scarani, Sofyan Iblisdir, Nicolas Gisin, and Antonio Acín. Quantum cloning. *Reviews of Modern Physics*, 77:1225–1256, Nov 2005.
- [3] Nicolas J Cerf, A Ipe, and X Rottenberg. Cloning of continuous quantum variables. *Physical Review Letters*, 85(8):1754, 2000.
- [4] Stephen M Barnett and Sarah Croke. Quantum state discrimination. *Advances in Optics and Photonics*, 1(2):238–278, 2009.
- [5] Yuxiang Yang, Giulio Chiribella, and Gerardo Adesso. Certifying quantumness: Benchmarks for the optimal processing of generalized coherent and squeezed states. *Physical Review A*, 90(4):042319, 2014.
- [6] Benjamin Schumacher. Quantum coding. *Physical Review A*, 51(4):2738, 1995.
- [7] Richard Jozsa and Benjamin Schumacher. A new proof of the quantum noiseless coding theorem. *Journal of Modern Optics*, 41(12):2343–2349, 1994.
- [8] Hoi-Kwong Lo. Quantum coding theorem for mixed states. *Optics Communications*, 119(5-6):552–556, 1995.
- [9] Si Wu, Shun-ichi Amari, and Hiroyuki Nakahara. Population coding and decoding in a neural field: a computational study. *Neural Computation*, 14(5):999–1026, 2002.
- [10] M. Plesch and V. Bužek. Efficient compression of quantum information. *Physical Review A*, 81(3):032317, 2010.
- [11] Lee A. Rozema, Dylan H. Mahler, Alex Hayat, Peter S. Turner, and Aephraim M. Steinberg. Quantum data compression of a qubit ensemble. *Physical Review Letters*, 113:160504, Oct 2014.
- [12] Yuxiang Yang, Giulio Chiribella, and Daniel Ebler. Efficient quantum compression for ensembles of identically prepared mixed states. *Physical Review Letters*, 116:080501, Feb 2016.
- [13] Yuxiang Yang, Giulio Chiribella, and Masahito Hayashi. Optimal compression for identically prepared qubit states. *Physical Review Letters*, 117:090502, Aug 2016.
- [14] Masahito Hayashi and Vincent Tan. Minimum rates of approximate sufficient statistics. *arXiv preprint arXiv: 1612.02542*, 2016.
- [15] Masahito Hayashi and Keiji Matsumoto. Asymptotic performance of optimal state estimation in qubit system. *Journal of Mathematical Physics*, 49(10):102101, 2008.
- [16] Mădălin Guță and Jonas Kahn. Local asymptotic normality for qubit states. *Physical Review A*, 73(5):052108, 2006.
- [17] Mădălin Guță and Anna Jenčová. Local asymptotic normality in quantum statistics. *Communications in Mathematical Physics*, 276(2):341–379, 2007.
- [18] Jonas Kahn and Mădălin Guță. Local asymptotic normality for finite dimensional quantum systems. *Communications in Mathematical Physics*, 289(2):597–652, 2009.
- [19] Masahito Hayashi. Asymptotic quantum estimation theory for the thermal states family. In *Quantum Communication, Computing, and Measurement 2*, pages 99–104. Springer, 2002.
- [20] Wataru Kumagai and Masahito Hayashi. Quantum hypothesis testing for gaussian states: Quantum analogues of  $\chi^2$ , t-, and f-tests. *Communications in Mathematical Physics*, 318(2):535–574, 2013.
- [21] Carlton M Caves. Quantum limits on noise in linear amplifiers. *Physical Review D*, 26(8):1817, 1982.
- [22] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- [23] Jeongwan Haah, Aram W. Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 913–925, New York, NY, USA, 2016. ACM.
- [24] Mădălin Guță, Peter Bowles, and Gerardo Adesso. Quantum-teleportation benchmarks for independent and identically distributed spin states and displaced thermal states. *Physical Review A*, 82(4):042310, 2010.
- [25] Alexander Semenovitch Holevo. Bounds for the quantity of information transmitted by a quantum communication channel. *Problemy Peredachi Informatsii*, 9(3):3–11, 1973.
- [26] Robert M Fano and David Hawkins. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29(11):793–794, 1961.
- [27] Michał Horodecki. Limits for compression of quantum information carried by ensembles of mixed states. *Physical Review A*, 57(5):3364, 1998.
- [28] Andreas Winter. Coding theorem and strong converse for quantum channels. *IEEE Transactions on information theory*, 45(7):2481, 1999.
- [29] Mark Wilde. The quantum data-compression theorem. In *Quantum information theory*, chapter 18. Cambridge University Press, 2013.
- [30] FAM De Oliveira, MS Kim, PL Knight, and V Buck. Properties of displaced number states. *Physical Review A*, 41(5):2645, 1990.

## APPENDIX

## A. Proof of Lemma 1.

For any  $x > 0$ , we define a series of  $t + 1 = \lfloor n^{1/2+x} \rfloor$  intervals  $L_0, \dots, L_t$  as

$$\begin{aligned}
 L_0 &= \{0\} \\
 L_i &= \{(i-1) \lfloor n^{(1-x)/2} \rfloor + 1, \dots, i \lfloor n^{(1-x)/2} \rfloor\} \quad 0 < i < t \\
 L_t &= \{(t-1) \lfloor n^{(1-x)/2} \rfloor + 1, \dots\}.
 \end{aligned}$$

For any non-negative integer  $m$ , we denote by  $i(m)$  the index of the interval containing  $m$ , i.e.  $m \in L_{i(m)}$ .

To design the compression protocol, we first notice that the  $n$ -fold thermal state can be written in the form

$$\left(\rho_{\beta}^{(\text{thm})}\right)^{\otimes n} = (1 - \beta)^n \sum_{\vec{m}} \beta^{|\vec{m}|} |\vec{m}\rangle \langle \vec{m}|, \quad (76)$$

where  $|\vec{m}\rangle = |m_1\rangle \otimes \cdots \otimes |m_n\rangle$  is the photon number basis of  $n$  modes and  $|\vec{m}| := m_1 + \cdots + m_n$ . The compression protocol runs as follows:

- *Encoder.* First perform projective measurement in the photon number basis of  $n$  modes, which yields an  $n$ -dimensional vector  $\vec{m}$ . Then compute  $i(|\vec{m}|)$  and encode it into a classical memory. The encoding channel can be represented as

$$\mathcal{E}_{n,x}^{(\text{thm})}(\rho) := \sum_{\vec{m}} \langle \vec{m} | \rho | \vec{m} \rangle |i(|\vec{m}|)\rangle \langle i(|\vec{m}|)|.$$

- *Decoder.* Read an integer  $i$  from the memory. If  $i \geq t$  prepare a fixed state  $|\vec{t}\rangle \langle \vec{t}|$  (defined below); if  $i < t$  perform random sampling in the interval  $L_i$ . For each outcome  $\hat{m}$  of the sampling, prepare the  $n$ -mode state

$$\binom{n + \hat{m} - 1}{\hat{m}}^{-1} \sum_{\vec{m}: |\vec{m}| = \hat{m}} |\vec{m}\rangle \langle \vec{m}|.$$

Then the decoding channel can be represented as

$$\mathcal{D}_{n,x}^{(\text{thm})}(|i\rangle \langle i|) := \begin{cases} \sum_{\vec{m}: |\vec{m}| \in L_i} \frac{|\vec{m}\rangle \langle \vec{m}|}{|L_i| \binom{n + |\vec{m}| - 1}{|\vec{m}|}} & i < t \\ |\vec{t}\rangle \langle \vec{t}| & i = t \end{cases}$$

$$\vec{t} = ((t-1) \lfloor n^{(1-x)/2} \rfloor + 1, \dots, (t-1) \lfloor n^{(1-x)/2} \rfloor + 1).$$

It is straightforward from definition that the protocol uses  $\log(t+1) = (1/2 + x) \log n + o(1)$  classical bits. What remains is to bound the error of the protocol. First, we notice that the recovered state is

$$\mathcal{D}_{n,x}^{(\text{thm})} \circ \mathcal{E}_{n,x}^{(\text{thm})} \left[ \left( \rho_{\beta}^{(\text{thm})} \right)^{\otimes n} \right] = \sum_{\vec{m}} (\rho_{\beta,n})_{\vec{m}} |\vec{m}\rangle \langle \vec{m}| \quad (77)$$

$$(\rho_{\beta,n})_{\vec{m}} = \begin{cases} (1 - \beta)^n \beta^{|\vec{m}|} \sum_{m \in L_{i(|\vec{m}|)}} \frac{\beta^{m - |\vec{m}|} \binom{n + m - 1}{m}}{|L_{i(|\vec{m}|)}| \binom{n + |\vec{m}| - 1}{|\vec{m}|}} & i(|\vec{m}|) < t \\ \sum_{\vec{m}': i(|\vec{m}'|) = t} (1 - \beta)^n \beta^{|\vec{m}'|} & \vec{m} = \vec{t} \\ 0 & \text{else.} \end{cases} \quad (78)$$

We choose  $S$  as the minimal set satisfying i)  $S$  is a union of several intervals chosen from the set  $\{L_i\}$  and ii)

$$S \supset \left[ \frac{\beta n}{1 - \beta} - n^{(1-x)/2}, \frac{\beta n}{1 - \beta} + n^{(1-x)/2} \right]. \quad (79)$$

Apparently,  $S \subset [0, n^{1+x/2}]$  for large enough  $n$ . Then the error can be bounded as

$$\begin{aligned} \epsilon_{\text{thm}} &= \frac{1}{2} \left\| \mathcal{D}_{n,x}^{(\text{thm})} \circ \mathcal{E}_{n,x}^{(\text{thm})} \left[ \left( \rho_{\beta}^{(\text{thm})} \right)^{\otimes n} \right] - \left( \rho_{\beta}^{(\text{thm})} \right)^{\otimes n} \right\|_1 \\ &= \frac{1}{2} \sum_{\vec{m}} \left| (1 - \beta)^n \beta^{|\vec{m}|} - (\rho_{\beta,n})_{\vec{m}} \right| \\ &\leq \sum_{\vec{m}: |\vec{m}| \notin S} (1 - \beta)^n \beta^{|\vec{m}|} + \frac{1}{2} \left[ \sum_{\vec{m}: |\vec{m}| \in S} (1 - \beta)^n \beta^{|\vec{m}|} \right] \max_{\substack{m, m' \in L_i \\ L_i \cap S \neq \emptyset}} \left| \sum_{m \in L_i} \frac{\beta^m \binom{n + m - 1}{m}}{|L_i| \beta^{m'} \binom{n + m' - 1}{m'}} - 1 \right| + \sum_{\vec{m}': i(|\vec{m}'|) = t} (1 - \beta)^n \beta^{|\vec{m}'|} \\ &\leq \sum_{\vec{m}: |\vec{m}| \notin S} (1 - \beta)^n \beta^{|\vec{m}|} + \frac{1}{2} \left[ \sum_{\vec{m}: |\vec{m}| \in S} (1 - \beta)^n \beta^{|\vec{m}|} \right] \max_{\substack{m, m' \in L_i \\ L_i \cap S \neq \emptyset}} \left| \frac{\beta^m \binom{n + m - 1}{m}}{\beta^{m'} \binom{n + m' - 1}{m'}} - 1 \right| + O((1 - \beta)^n \beta^n) \\ &\leq \sum_{\vec{m}: |\vec{m}| \notin S} (1 - \beta)^n \beta^{|\vec{m}|} + \frac{1}{2} \max_{\substack{m, m' \in L_i \\ L_i \cap S \neq \emptyset}} \left| \frac{\beta^m \binom{n + m - 1}{m}}{\beta^{m'} \binom{n + m' - 1}{m'}} - 1 \right| + O((1 - \beta)^n \beta^n). \end{aligned}$$

On one hand, we notice that  $|\vec{m}|$  is the sum of  $n$  i.i.d. random variables with geometric distribution  $\{(1-\beta)\beta^i\}_{i=0}^\infty$  and thus, by Central Limit Theorem, the first error term scales as

$$\sum_{\vec{m}: |\vec{m}| \notin S} (1-\beta)^n \beta^{|\vec{m}|} = O \left[ \operatorname{erfc} \left( \frac{n^{x/2}(1-\beta)}{\sqrt{2\beta}} \right) \right] = e^{-\Omega(n^x)}$$

where  $\operatorname{erfc}(x) := (2/\pi) \int_x^\infty e^{-s^2} ds$  is the complementary error function. On the other hand, in the second error term,  $m$  and  $m'$  are in the same order as  $n$ , so the second term can be bounded as

$$\begin{aligned} \max_{\substack{m, m' \in L_i \\ L_i \cap S \neq \emptyset}} \left| \frac{\beta^m \binom{n+m-1}{m}}{\beta^{m'} \binom{n+m'-1}{m'}} - 1 \right| &= \max_{\substack{m, m' \in L_i \\ L_i \cap S \neq \emptyset}} \left| \beta^{m-m'} \frac{(n+m-1) \cdots (n+m')}{m \cdots (m'+1)} - 1 \right| \\ &= \max_{\substack{m, m' \in L_i \\ L_i \cap S \neq \emptyset}} \left| \left( \frac{\beta m + \beta n}{m} \right)^{m-m'} \left[ 1 + O \left( \frac{|L_i|^2}{n} \right) \right] - 1 \right| \\ &= | [1 + O(n^{-x})] [1 + O(n^{-x})] - 1 | \\ &= O(n^{-x}). \end{aligned}$$

Therefore, we have proved Eq. (9).

### B. Proof of Lemma 2.

For any input state  $\rho$  we have

$$\varepsilon(\rho) \leq \frac{1}{2} \|P_N \rho P_N - \rho\|_1 + \frac{1}{2} [1 - \operatorname{Tr}(\rho P_N)] \quad (80)$$

$$\leq \frac{1}{2} \left[ 2\sqrt{1 - \operatorname{Tr}(\rho P_N)} + 1 - \operatorname{Tr}(\rho P_N) \right] \quad (81)$$

$$\leq \frac{3}{2} \sqrt{1 - \operatorname{Tr}(\rho P_N)}. \quad (82)$$

The second inequality came from the gentle measurement lemma [28], [29]. Substituting  $\rho_{\alpha, \beta} = D_\alpha \rho_\beta^{(\text{thm})} D_\alpha^\dagger$  into the above bound, we express the truncation error for  $\rho_{\alpha, \beta}$  as

$$\varepsilon(\rho_{\alpha, \beta}) \leq \frac{3}{2} \sqrt{1 - \operatorname{Tr} \left[ D_\alpha \rho_\beta^{(\text{thm})} D_\alpha^\dagger P_N \right]}. \quad (83)$$

We now bound the trace part in the right hand side of the last inequality as

$$\begin{aligned} 1 - \operatorname{Tr} \left[ D_\alpha \rho_\beta^{(\text{thm})} D_\alpha^\dagger P_N \right] &= \sum_{k \leq N} \sum_{l=0}^\infty (1-\beta) \beta^l |\langle l | D_\alpha | k \rangle|^2 \\ &\leq \max_{l \leq l_0} \sum_{k > N} |\langle l | D_\alpha | k \rangle|^2 + \sum_{l \geq l_0} (1-\beta) \beta^l \\ &\leq \max_{l \leq l_0} \sum_{k > N} |\langle l | D_\alpha | k \rangle|^2 + \beta^{l_0} \end{aligned} \quad (84)$$

Here we set  $l_0 = N^{x/8}$ . Notice that  $|\langle l | D_\alpha | k \rangle|^2$  is the photon number distribution of a displaced number state [30], which can be bounded as

$$\begin{aligned} |\langle l | D_\alpha | k \rangle|^2 &= \frac{e^{-|\alpha|^2} (|\alpha|^2)^{k+l}}{k! l!} \left| \sum_{i=0}^{\min\{k, l\}} \frac{k! l! (-|\alpha|^2)^{-i}}{i! (k-i)! (l-i)!} \right|^2 \\ &\leq \frac{e^{-|\alpha|^2} (|\alpha|^2)^{k+l}}{k! l!} \left| \sum_{i=0}^k \binom{k}{i} \left( \frac{l}{|\alpha|^2} \right)^i \right|^2 \\ &= \frac{e^{-|\alpha|^2} |\alpha|^{2l}}{k! l!} \left( \frac{l + |\alpha|^2}{|\alpha|} \right)^{2k}. \end{aligned}$$

Then we can bound the first term in (84) as

$$\begin{aligned} \max_{l \leq l_0} \sum_{k > N} |\langle l | D_\alpha | k \rangle|^2 &\leq \max_{l \leq l_0} \frac{|\alpha|^{2l}}{l!} \sum_{k > N} \frac{e^{-|\alpha|^2}}{k!} \left( \frac{l + |\alpha|^2}{|\alpha|} \right)^{2k} \\ &= \max_{l \leq l_0} \frac{e^{2l + \frac{l^2}{|\alpha|^2}} |\alpha|^{2l}}{l!} \sum_{k > N} \mathbf{Pois}_{\lambda_\alpha}(k), \end{aligned}$$

where  $\mathbf{Pois}_{\lambda_\alpha}(k)$  is the Poisson distribution with mean  $\lambda_\alpha = (l + |\alpha|^2)^2 / |\alpha|^2$ . Notice that  $[\lambda_\alpha - N^{1/2+x/2}, \lambda_\alpha + N^{1/2+x/2}] \subseteq [0, N]$  and thus we have

$$\begin{aligned} \max_{l \leq l_0} \sum_{k > N} |\langle l | D_\alpha | k \rangle|^2 &\leq \max_{l \leq l_0} \frac{|\alpha|^{2l} e^{2l + \frac{l^2}{|\alpha|^2}}}{l!} \sum_{|k - \lambda_\alpha| > N^{1/2+x/2}} \mathbf{Pois}_{\lambda_\alpha}(k) \\ &= \max_{l \leq |\alpha|^{x/4}} \frac{|\alpha|^{2l} e^{2l + \frac{l^2}{|\alpha|^2}}}{l!} e^{-\Omega(N^{x/2})} \\ &= e^{-\Omega(N^{x/4})}. \end{aligned}$$

having used the tail bound for Poisson distribution and  $l_0 = N^{x/8}$ . Substituting the above bound into Eq. (84), we get

$$\sqrt{1 - \text{Tr} [D_\alpha \rho_\beta^{(\text{thm})} D_\alpha^\dagger P_\alpha]} \leq \beta^{\Omega(N^{x/8})} + e^{-\Omega(N^{x/4})}.$$

Finally, substituting the above inequality into Eq. (83), we can bound the error as

$$\varepsilon(\rho_{\alpha, \beta}) = \beta^{\Omega(N^{x/8})} + e^{-\Omega(N^{x/4})}. \quad (85)$$